A General Jackknife within Each Stratum Variance Estimator

Mousa. A. M., El Sayed S. M., Abdel Latif S. H.

Abstract— **In survey sampling, accuracy of point estimates are assessed using variance estimates. Variance estimates becomes difficult when we have non-linear point estimators or complex sampling designs. The jackknife has been suggested as a useful cutoff resampling technique to overcome these difficulties. The approximation of sampling relative error for the total in stratified sampling without replacement from finite populations will be reviewed in this paper. And a general jackknife within each stratum estimator of Cao, R. et al (2013) is derived.**

Index Terms— **Jackknife; Relative Error; Sampling Survey; Stratified Sampling; Variance Estimation.**

I. INTRODUCTION

Hansen and Hurwitz(1943) gave the first theoretical framework of unequal probability sampling with replacement. Also, Madow (1949) introduced the concept of unequal probability sampling without replacement. Horvitz and Thompson (HT) (1952) developed the theoretical ideas of unequal probability sampling without replacement and proposed an estimator of the population total. Then, they obtained an estimator of the actual variance of their estimator of the total and an alternative expression of HT estimator developed by Yates and Grundy(1953). The previous estimators had some complexities in computing.

To overcome these computing problems, Hartley and Rao(1962) proposed approximate expression for HT(1952) estimator using random systematic sampling procedure. Also, Basit and Shahbaz(2006), Shahbaz and Hanif,(2007), Alodat(2009), and Razvi and Khan(2013). Hassan et al.(2009) empirically compared various approximate formulae for variance of HT estimator.

According to the resampling techniques, Quenouille (1949) introduced the jackknife as a method of reducing the bias of an estimator of a serial correlation coefficient. In a 1956, Quenouille generalized its method.

Variance estimates becomes difficult when we have non-linear point estimators and complex sampling designs. Durbin(1959) applied Quenouille's method, in sample survey context, in finite population to study its use in ratio estimation, using $g = 2$ groups. Rao(1965) and Rao and Webster(1966) studied the optimal choice of g for bias reduction in ratio estimation, and showed that $g = n$ is the optimal choice.

With stratified sampling design using jackknife method under unequal probability sampling, Lee(1973) attempted to develop an efficient scheme for variance estimation. Campbell(1980) proposed a generalized jackknife variance estimator and gave a general definition of a jackknife pseudo value that is applicable for unequal probability sampling and stratification, and generalized the jackknife variance estimate so that it applied to any sample design for which the variance of an estimated mean can be estimated (exist).

Therefore, the resampling techniques, in the context of sampling survey, has been widely studied and developed to handle stratified multistage sampling by Jones(1974), Kish and Frankel(1974), Krewski and Rao(1981), and Kovar et al.(1988), and the properties of various forms of the jackknife estimator for this case have been studied theoretically and empirically.

Berger and Skinner(2005) established the consistency of Campbell’s generalized jackknife variance estimator. They also compared the performance of Campbell's jackknife in a single stage context with standard single stage jackknife such as in Kish and Frankel(1974). Also, Berger(2007) modified Campbell's estimator by proposing a simple jackknife variance estimator. Berger's estimator is consistent under unistage stratified sampling without replacement. Escobar and Berger(2011) proposed two generalized jackknife variance estimators suitable for functions of HT(1952) point estimators.

Rao(2009) presented a brief overview of early uses of re-sampling methods in survey sampling, and provided an appraisal of more recent re-sampling methods for variance estimation and inference for small areas.

While, Cao, et al.(2013), studied the problem of approximating the sampling relative error of point estimates derived from large sample surveys on a finite population using stratified random sampling design without replacement. They proposed three jackknife methods and compared it with the plug-in and two bootstrap techniques. The first one(J1) considered by removing a sample value at each iteration, the second one(J2) constructed by removing a stratum at each iteration, and the third estimate(J3) constructed by considering the variance of $\hat{\theta}$ as a linear combination of variances of statistics constructed at stratum level, and these variances are previously estimated by jackknife in each stratum. They examined and compared the different procedures by extensive simulation study.

Amany Mousa Mohamed Mousa, Applied Statistics and Econometrics, Cairo University/ Institute of Statistical Studies and Research, Giza, Egypt, 01222412186.

Sayed Mesheal El Sayed, Applied Statistics and Econometrics, Cairo University/ Institute of Statistical Studies and Research, Giza, Egypt, 01001116532.

Abdel Latif, S. H, Applied Statistics and Econometrics, Cairo University/ Institute of Statistical Studies and Research, Giza, Egypt, Phone/ 01008233092

II. THE SAMPLING RELATIVE ERROR

The sampling error of $\hat{\theta}$ can be presented in relative terms, using the variation coefficient of the estimator given by:

$$E_{rel}(\hat{\theta}) = \sqrt{\frac{var(\hat{\theta})}{E(\hat{\theta})}} \qquad (1)$$

I. THE JACKKNIFE METHOD

A. The Grouped Quenouille-Tukey (QT) Jackknife Method

The sample of size n independent and identically distributed(iid) observations $x_1, x_2, \dots, x_n$ is divided into g non-overlapping groups $G_1, G_2, \dots, G_g$, each of size d , assuming that $n = dg$ with the i^{th} jackknife sample $S_{(i)}$=($S_1, S_2, \dots, S_{(i-1)d}, S_{id+1}, S_{id+2}, \dots, S_{gd}$) and i^{th} group ($S_{(i-1)d+1}, S_{(i-1)d+2}, \dots, S_{id}$) are deleted in turn and the "delete-group" estimates $\hat{\theta}_i$, $i = 1, 2, \dots g$, are computed, where $\hat{\theta}_i$ denotes the estimator of θ based on the sample of size $n - d = d(g-1)$, which are named pseudo estimates . Quenouille showed that the estimator

$\hat{\theta}_J = \sum_{i=1}^{g} \tilde{\theta}_i / g$,

where, $\tilde{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_i$ and these are named pseudo values.

$$\hat{\theta}_J = \frac{1}{g}\sum_{i=1}^{g}\tilde{\theta}_i, \qquad (2)$$

$\hat{\theta}_J$ can be expressed in terms of the pseudo estimates as

$\hat{\theta}_J = g\hat{\theta} - \frac{g-1}{g}\sum_{i=1}^{g}\hat{\theta}_i$,

or even,

$$\hat{\theta}_J = \hat{\theta} + (g-1)(\hat{\theta} - \hat{\theta}_{(.)}), \qquad (3)$$

with $\hat{\theta}_{(.)} = \frac{1}{g}\sum_{i=1}^{g}\hat{\theta}_i$.

In an abstract, Tukey(1958) noted that for $g = n$ and $\hat{\theta} = \bar{x}$, the sample mean, the" pseudo-values" $\tilde{\theta}_i$ reduce to $\tilde{\theta}_i = x_i$ and hence iid . Motivated by this result, Tukey suggested regarding the $\tilde{\theta}_i$ as iid for general $\hat{\theta}$ and then using

$$v(\hat{\theta}_J) = \sum_{i=1}^{n}(\tilde{\theta}_i - \hat{\theta}_J)^2 / n(n-1). \qquad (4)$$

$$= \frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_{(.)})^2, \qquad (5)$$

$\hat{\theta}_{(.)} = \sum_{i=1}^{n}\hat{\theta}_i / n$. As the "jackknife" variance estimator of $\hat{\theta}_J$ or $\hat{\theta}$.

B. The Cao et al. Jackknife Method

As we mentioned previously, Cao et al.(2013) proposed three jackknife estimators for the variance of $\hat{\tau}$.

Case 1: Jackknife Leaving One Sample Value Out:

The jackknife estimator for the relative error is:

$$\hat{\epsilon}_J(\hat{\tau}) = \frac{1}{\hat{\tau}}\left(\frac{n-1}{n}\sum_{r=1}^{L}\frac{N_r{}^2}{(n_r-d)}(S_r^2)\right)^{1/2} , \qquad (6)$$

where, L and N_i will denote the number of strata and the population size of the i^{th} stratum, $i = 1, \dots, L$, respectively. And $\{x_{ij},\ j = 1, \dots, n_i,\ i = 1, \dots, L\}$ be a stratified random sample without replacement of the variable of interest X, of size $n = \sum_{i=1}^{L} n_i$, n_i being the sample size within the i^{th} stratum. The unbiased estimator of the total,$\hat{\tau}$, and the sample variances,S_i^2, are respectively given by;

$\hat{\tau} = N\hat{\mu} = \sum_{i=1}^{L} N_i \bar{x}_{i.} = \sum_{i=1}^{L} F_i n_i \bar{x}_{i.} = \sum_{i=1}^{L} F_i x_{i.}$

$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2$,

where, $\hat{\mu}$ the unbiased estimator of population mean μ, is

$\hat{\mu} = \sum_{i=1}^{L}\frac{N_i}{N}\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} = \frac{1}{N}\sum_{i=1}^{L} F_i n_i \bar{x}_{i.}$

$= \frac{1}{N}\sum_{i=1}^{L} F_i x_{i.}$

and, $F_i = N_i / n_i$ the elevation factors of each stratum. Being $\bar{x}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ and $x_{i.} = \sum_{j=1}^{n_i} x_{ij}$.

i.e., the population mean is given by

$\mu = \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{N_i} X_{ij} = \sum_{i=1}^{L}\frac{N_i}{N}\bar{X}_{i.}$,

and the population total is given by

$\tau = \sum_{i=1}^{L}\sum_{j=1}^{N_i} X_{ij} = N\mu = \sum_{i=1}^{L} N_i \bar{X}_{i.} = \sum_{i=1}^{L} X_{i.}$,

where, X_{ij} is the j^{th} element of the i^{th} population stratum, with $j = 1, \dots, N_i$, $i = 1, \dots, L$. Being $\bar{X}_{i.} = \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij}$ and $X_{i.} = \sum_{j=1}^{N_i} X_{ij}$.

Case 2: Jackknife Leaving One Stratum Out

Two variants of the jackknife estimator were introduced by considering different ways of averaging the pseudo values. First, weighted mean was used, where each pseudo value was weighted by the population size of some strata removed in calculation. So, the jackknife estimator for the relative error

$$\hat{\epsilon}_{J,1r}(\hat{\tau}) = \frac{1}{\hat{\tau}}\left(\widehat{var}_{J,1r}(\hat{\tau})\right)^{1/2} , \qquad (7)$$

$\widehat{var}_{J,1r}(\hat{\tau}) = \sum_{r=1}^{L}\frac{N_r}{N}\left(\frac{N-N_r}{N}\right)\left(\hat{\tau}_{(r)} - \hat{\tau}_{(.)}^{1r}\right)^2$,

$\widehat{var}_{J,1r}(\hat{\tau}) = \sum_{r=1}^{L}\frac{N_r}{N}\left(\frac{N-N_r}{N}\right)\left[\frac{N(\hat{\tau} - F_r x_{r.})}{N-N_r} - \sum_{i=1}^{L}\frac{N_i(\hat{\tau} + F_i x_{i.})}{N-N_i}\right]^2$.

An alternative variant of the jackknife leaving some strata out was obtained when all the strata contribute with the same weight in the estimation, and the jackknife estimator for the relative error is

$$\hat{\epsilon}_{J,2r}(\hat{\tau}) = \frac{1}{\hat{\tau}}\left(\widehat{var}_{J,2r}(\hat{\tau})\right)^{1/2} . \qquad (8)$$

$\widehat{var}_{J,2r}(\hat{\tau}) = \frac{L-1}{L}\sum_{r=1}^{L}\left(\hat{\tau}_{(r)} - \hat{\tau}_{(.)}^{2r}\right)^2$,

$= \frac{L-1}{L}\sum_{r=1}^{L}\left[\frac{N(\hat{\tau} - F_r x_{r.})}{N-N_r} - \frac{N}{L}\sum_{i=1}^{L}\frac{\hat{\tau} + F_i x_{i.}}{N-N_i}\right]^2$

Case 3: Jackknife within Each Stratum

The jackknife estimator for the relative error is

$$\hat{\epsilon}_{J,3}(\hat{\tau}) = \frac{1}{\hat{\tau}}\left(\widehat{var}_{J,3}(\hat{\tau})\right)^{1/2} , \qquad (9)$$

$\widehat{var}_{J,3}(\hat{\tau}) = \sum_{i=1}^{L}\frac{N_i^2}{n_i}S_i^2$.

El Sayed and Abdel Latif (2015) introduced a general jackknife method of Cao, R. et al (2013) for(6), the jackknife estimator for the relative error in case(1) of leaving one sample value out, by removing a group of values which is given by the following case;

Case 4: Jackknife Leaving One Sample Group Out

The jackknife estimator of $var(\hat{\tau})$ is given by

$\widehat{var}_{J,d}(\hat{\tau}) = \frac{g-1}{g}\sum_{r=1}^{L}\frac{N_r{}^2}{(n_r-d)}(S_r^2)$.

And the derived jackknife estimator for the relative error in the case of leaving one sample group out is

$$\hat{\epsilon}_J(\hat{\tau}) = \frac{1}{\hat{\tau}}\left(\frac{g-1}{g}\sum_{r=1}^{L}\frac{N_r{}^2}{(n_r-d)}(S_r^2)\right)^{1/2} , \qquad (10)$$

its proof exist in El Sayed and Abdel Latif(2015).

III. THE GENERAL JACKKNIFE WITHIN EACH STRATUM

Here, we proposed a general jackknife method of Cao, R. et al.(2013) for(9), the jackknife estimator for the relative error in case(3), the jackknife estimator within each stratum. According to $\hat{\tau} = N\hat{\mu} = \sum_{i=1}^{L} N_i \bar{x}_{i.}$, the variance of $\hat{\tau}$ can be expressed as a linear combination of the variances of the sample means within each stratum

$$var(\hat{\tau}) = \sum_{i=1}^{L} N_i^2 var(\bar{x}_{i.}) \qquad (11)$$

Hence, a general jackknife approximation to the variance of $\hat{\tau}$ can be obtained by estimating each $var(\bar{x}_{i.})$ with the jackknife method (when its pseudovalue obtained by eliminating group of observations of the r^{th} stratum) and replacing these estimators in(11). For the jackknife estimator of $var(\bar{x}_{i.})$, the pseudovalues are defined as

$$u_i^{(G)} = \bar{y}_{i.}^{(G)}$$
$$= \frac{1}{g_i-1}\sum_{\substack{j=1\\ j\neq G}}^{g_i} y_{ij} = \frac{1}{g_i-1} y_{i.}^{(G)} ,$$

where, y_{ij} is the j^{th} group of the i^{th} stratum.

For $G = 1, 2, \dots, g_i$, and their mean is given by

$$u_i^{(.)} = \frac{1}{g_i}\sum_{G=1}^{g_i} u_i^{(G)}$$
$$= \frac{1}{g_i(g_i-1)}\sum_{G=1}^{g_i}\sum_{\substack{j=1\\ j\neq G}}^{g_i} y_{ij}$$
$$= \frac{1}{g_i(g_i-1)}\sum_{j=1}^{g_i}(g_i - 1)y_{ij}$$
$$= \bar{y}_{i.}$$

Then, the jackknife estimator of the variance of the sample mean of the i^{th} stratum is

$$\widehat{var}_J(\bar{y}_{i.}) = \frac{g_i-1}{g_i}\sum_{G=1}^{g_i}\left(u_i^{(G)} - u_i^{(.)}\right)^2$$
$$= \frac{g_i-1}{g_i}\sum_{G=1}^{g_i}\left[\frac{\sum_{\substack{j=1,\\ j\neq G}}^{g_i} y_{ij}}{g_i} - \frac{\sum_{j=1}^{g_i} y_{ij}}{g_i}\right]^2$$
$$= \frac{g_i-1}{g_i}\sum_{G=1}^{g_i}\left[\sum_{j=1}^{g_i}\frac{y_{ij}}{g_i-1} - \sum_{j=1}^{g_i}\frac{y_{ij}}{g_i} - \frac{y_{iG}}{g_i-1}\right]^2$$
$$= \frac{g_i-1}{g_i}\sum_{G=1}^{g_i}\left[\frac{\sum_{j=1}^{g_i} y_{ij}}{g_i(g_i-1)} - \frac{y_{iG}}{g_i-1}\right]^2$$
$$= \frac{\left(\sum_{j=1}^{g_i} y_{ij}\right)^2}{g_i^2(g_i-1)} + \frac{\sum_{G=1}^{g_i} y_{iG}^2}{g_i(g_i-1)} - \frac{2\left(\sum_{j=1}^{g_i} y_{ij}\right)^2}{g_i^2(g_i-1)}$$
$$= \frac{1}{g_i(g_i-1)}\sum_{G=1}^{g_i} y_{iG}^2 - \frac{\left(\sum_{j=1}^{g_i} y_{ij}\right)^2}{g_i^2(g_i-1)}$$
$$= \frac{1}{(g_i - 1)g_i}\sum_{j=1}^{g_i}\left(y_{ij} - \bar{y}_{i.}\right)^2$$
$$= \frac{s_i^2}{g_i},$$

By using the previous jackknife estimations the following estimator will be obtained,

$$\widehat{var}_{J,3}(\hat{\tau}) = \sum_{i=1}^{L}\frac{N_i^2}{g_i} s_i^2$$

The corresponding jackknife estimation of the relative error is given by

$$\hat{\epsilon}_{J,3}(\hat{\tau}) = \frac{1}{\hat{\tau}}\left(\widehat{var}_{J,3}(\hat{\tau})\right)^{1/2}$$

REFERENCES

[1] Alodat, N. A., On unequal probability sampling without replacement sample size 2, *Int. J. Open Problems Comp. Math.* 2 (2009), 108-112.

[2] Basit, A. and Shahbaz, M. Q., A class of selection procedure for unequal probability sampling without replacement and a sample of size 2, *J. Statist.* 13 (2006), 26-32.

[3] Berger, Y. G., A jackknife variance estimator for unistage stratified samples with unequal probabilities, *Biometrika* 94 (2007), 953-964.

[4] Berger, Y. G. and Skinner, C. J., A jackknife variance estimator for unequal probability sampling, *J. Roy. Statist. Soc. Ser. B* 67 (2005), 79-89.

[5] Cao, R., Vilar, J. A., Vilar, J. M., and López, A. K., Sampling Error Estimation in Stratified Surveys, *Open Journal of Statistics* 3(2013), 200-212.

[6] Campbell, C., A different view of finite population estimation, *Proc. Surv. Res. Meth. Sect. Am. Statist. Assoc*. (1980), 319-324.

[7] Durbin, J., A note on the application of Quenouille's method of bias reduction to the estimation of ratios, *Biometrika* 46(1959), 477-480.

[8] El Sayed, S. M. and Abdel Latif, S. H., General jackknife variance estimator for stratified sampling survey, *International Mathematical Forum* 10(2015), 377-383.

[9] Escobar, E. L. and Berger, Y. G., Jackknife variance estimation for functions of Horvitz & Thompson estimators under unequal probability sampling without replacement, *Int. Statistical Inst.: Proc. 58th World Statistical Congress*. Dublin (Session CPS046) (2011), 5031-5034.

[10] Hansen, M. H. and Hurwitz, W. N., On the theory of sampling from finite populations, *Ann. Math. Stat*. 14(1943), 333-362.

[11] Hartley, H. O. and Rao, J. N. K., Sampling with unequal probabilities and without replacement, *Ann. Math. Stat.* 33(1962), 350-374.

[12] Hassan, Y., Shahbaz, M. Q., and Hanif, M., Empirical comparison of some approximate variance formulae of Horvitz-Thompson estimator, *World Applied Sciences Journal* 5(2009), 597-599.

[13] Horvitz, D. G., and Thompson, D. J., A generalization of sampling without replacement from a finite universe, *J. Amer. Stat. Assoc.* 47(1952), 663-685.

[14] Jones, H. L., Jackknife estimation of functions of stratum means, *Biometrika* 61 (1974), 343-348.

[15] Kish, L. and Frankel, M. R., Inference from complex samples (with discussion), *J. Roy. Statist. Soc. Ser. B*. 36(1974), 1-37.

[16] Kovar, J. G., Rao, J. N. K. and Wu, C. F. J., Bootstrap and other methods to measure errors in survey estimates, *Can. J. Statist*. 16 (1988), 25-45.

[17] Krewski, D. and Rao, J. N. K., Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods, *Ann. Statist*. 9 (1981), 1010-1019.

[18] Lee, K., Variance estimation in stratified sampling, *J. Am. Statist. Assoc.* 68 (1973), 336-342.

[19] Madow, W. G., On the theory of systematic samplingII, *Ann. Math. Stat*. 20(1949), 333-354.

[20] Quenouille, M. H., Approximate tests of correlation in time-series, *J. Roy. Statist. Soc. Ser. B* 11(1949), 68-84.

[21] Quenouille, M. H., Notes on bias in estimation, *Biometrika* 34(1956), 353-360.

[22] Rao, J. N. K., A note on the estimation of ratios by Quenouille's method, *Biometrika* 52 (1965), 647-649.

[23] Rao, J. N. K., Jackknife and bootstrap methods for variance estimation from sample survey data, *Int. J. of Stat. Sci.* 9(Spcial Issue): 59-70.

[24] Rao, J. N. K. and Webster, J. T., On two methods for bias reduction, *Biometrika* 53(1966), 571-577.

[25] Razvi, N. A. and Khan, M.A., Sampling procedure with unequal probabilities of selection an generalization for fixed sample size, *International Journal of Statistika and Mathematika* 7 (2013), 12-13.

[26] Shahbaz, M. Q. and Hanif, M., A new approximate formula for variance of Horvitz-Thompson estimator using first order inclusion probabilities, *Pak .J .Stat. Oper. Res.* III(2007), 59-62.

[27] Tukey, J. W., Bias and confidence in not-quite large samples (Abstract), 1958.

[28] Yates, F. and Grundy, P.M., Selection without replacement from within strata with probability proportional to size, *J. Roy. Stat. Soc. Ser. B* 15(1953),253-261.

Amany Mousa Mohamed Mousa, Applied Statistics and Econometrics, Cairo University/ Institute of Statistical Studies and Research, Giza, Egypt, 01222412186.

Sayed Mesheal El Sayed, Applied Statistics and Econometrics, Cairo University/ Institute of Statistical Studies and Research, Giza, Egypt, 01001116532.

Abdel Latif, S. H, Applied Statistics and Econometrics, Cairo University/ Institute of Statistical Studies and Research, Giza, Egypt, Phone/ 01008233092